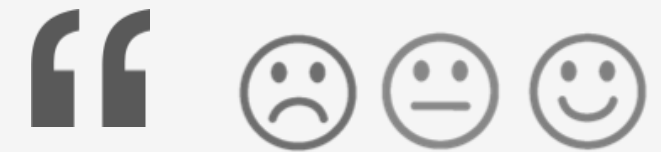


NLP 감정분석 기반
마케팅 시장 분석

키워드에 따른 커뮤니티 별 온라인 여론 감정 분석 시스템

빅데이터미네이터



커뮤니티별 이슈 키워드 감성 분석 시스템

각 커뮤니티의 제목, 본문, 댓글에 대한 크롤링을 진행하고, 이에 따른 주요 이슈 키워드를 추출한다.

또한 추출한 데이터를 바탕으로 감성 분석하여 결과를 도출하고자 한다.

커뮤니티별 익일 이슈 키워드와 그에 대한 감성 분석 결과를 사용자에게 알려주는 시스템이다.

1. Input

- 커뮤니티

: 네이트판, FM, 클리앙
(+DC, 뽀뿌)

- 크롤링 범위

→ 당일의 이슈 키워드 추출
(ex. 하루 중 일정 시간에 내용을
업데이트하여 사용자에게 알림)

: 제목, 본문, 댓글

2. 전처리

- 영어, 특수문자 등 제거
- 초성 (ㅋㅋ, ㅎㅎ 등) 제거

- 형태소 단위로 토큰화

→ 문장 단위로 순서는 유지해야함
(앞뒤 단어와의 연관성에 따른
감성 분석 가능)

- csv 파일로 저장 (프로토타입)

3. 프로토타입 구조

크롤링부 (+ 전처리)

↓
csv 파일 (추후 DB 구축)

↓ 감성어 사전
 (범용 사전 + 감성어 추가)

감성 분석기 => 결과 도출

가장 먼저 커뮤니티 선정한 커뮤니티인
네이트판, FM 코리아, 클리앙에 대한 크롤러를 만들었다.

제목, 본문, 댓글 내용에 대해 형태소 단위로 정리 한 뒤,
전체 형태소 개수, 명사 개수,
그리고 가장 많이 반복되는 키워드들을 뽑아보는 방식으로 진행했다.

커뮤니티 별 크롤링

01. 네이트판

- 당일 실시간 인기 톱 글이 올라오는 “툰커들의 선택 명예의 전당” 게시판의 제목, 본문, 댓글을 크롤링하였다.
- 제목 및 URL을 먼저 크롤링하고, 그 뒤 개별 본문으로 접근하여 각 본문 글과 댓글을 크롤링하는 순서로 진행했다.
- 크롤링한 데이터에 띄어쓰기가 올바르게 안 된 경우가 대다수여서 특수문자 등을 제거한 뒤, 한국어 전처리 패키지인 pykospacing을 이용하여 전체적으로 다시 띄어쓰기를 진행했다.
- KoNLPy 코엔엘파이, 한국어 형태소 분석기 내부의 Okt, Komoran, Twitter 등을 활용해 형태소 단위, 명사 단위의 분석을 진행하였다.

⇒ 앞으로 수정해 나가야하는 것

- 1) 전처리 단계에서 더 매끄러운 띄어쓰기 방식을 찾아볼 것
- 2) 형태소 분석기 사전에 온라인상 자주 사용되는 명사 등을 추가하여 전처리 단계에서 제외되지 않도록 해야 한다.
- 3) 더 의미 있는 최다 빈도 단어 추출을 위해 글자수 제한을 거는 등의 방법을 조사해 봐야한다.
- 4) 차단을 막기 위해 현재 sleep을 사용하고 있는데, 더 좋은 방법의 우회법을 찾아 봐야 한다.



전체 형태소 개수:
300613
전체 명사 개수:
35287

	단어	빈도수
0	남편	242
1	결혼	184
2	부모	156
3	엄마	153
4	여자	95
5	친구	81
6	남자	76
7	아이	74
8	본인	58
9	가요	56

[illegible][illegible]

코드 일부

커뮤니티 크롤링

02. FM



← → ↺ fmkorea.com

[보안 시스템에 의한 자동 차단]

사람이 아닌 자동화 프로그램에 의한 비정상적인 반복적인 접속이 탐지되어 에펠키리아 보안 시스템에 의해 사용하시는 IP가 차단되었습니다.
일부 브라우저에서 새로고침 키를 실수로 계속 누르고 있으면 차단될 수 있습니다.
공용 IP인 경우에는 다른 사용자의 행동에 의해 잘못된 차단도 가능합니다.
잘못 차단되었다고 생각하는 경우 VPN이나 warning 우회용 프로그램 이용하고 있는 경우 껴보시길 바랍니다.
브라우저 아닌 프로그램/앱으로 접속하고 있는 경우 해당 프로그램/앱 문제일 것 입니다.
관련 문제되는 앱이 의심되면 이메일로 사용하던 앱 이름을 알려주시길 바랍니다.

24 시간 이후에 자동으로 차단이 풀립니다.

잘못 차단된 경우 help@fmkorea.com에 하단의 정보와 함께 관련 문의를 하시길 바랍니다.

시간: 2021-05-16 02:11
IP: 182.219.121.18
나라: KR
접속 종류: 유선
ASNorg: LG POWERCOMM

	B	C	D	E	F	G	H	I	J	K	L	M	
	제목	본문	댓글										
0	이후리 불	머니게임	머니게임 이후리 말구요 아니근데 냉정하게 참가자들이 이 정도로 욕 먹어야 할거기 안에서 구라든 배신이든										
1	한자 진짜	수 천년 앞수 천년 앞서 간 선조들의 지혜 개인적으로 고 대 중국인들은 동양 문명의 시 조 맞다고 생각함 훌륭하신 분들											
2	유민상 맞	그냥 자 다그냥 자 다 일어났을 뿐인데 얼굴 존 나 시 커밍누가 봐도 정상적인 얼굴색 아님 올 고모부 당뇨 합병증으로 볼											
3	현대자동차	사무직 연	사무직 연구직도 공채 폐지 생산직도 모집 중단 원흥이 누구 다그건 노조라는 이름의 석파이프 든 빨간 조끼										
4	AV배우 하	미카미 유	미카미 유아 수개월 동안 연락 끊음 지금도 받아들이긴 힘들지만 미카미 유아가 행복하면 그걸로 됐다고 하심										
5	충격 현역	술치 다른	술치 다른 곳도 아니고 계룡대에서 이 정도로 밥을 주면 다른 곳은 얼마나 못 먹는 거야 연동이 아니라 진짜 군										
6	니가 100	을디 나이	을디 나이 두 자리인 새파란 눈들이 시골 경로당 가면 그런 거 없다 자식 잘 키운 눈이 최고 다 잘 살고 직업 중										
7	서양에서	성을 하나	성을 하나로 통일해서 한 가족으로 인정받고 유대감 형성을 위해 남편성으로 배운다고 함 일단 근본적으로는										
8	다시 봐도	정부가 직	정부가 직접 국민들에게 선택적 이 아닌 강제적으로 반일 감정 주입하고 외교적 무능함을 도재팬 해 줘' 로 시										
9	파이가 나	이제와서	이제 와서 파이가 왜 이런 이야기를 했는지 대응 이해가 간 다 파이의 생각은 가 오가 이 보내서 확실히 여자										
10	머니게임	일단 나	일단 나 일주일 내내 육해권 하는 것 보면 광기정병을 보는 것 같음모탈이런포 텐글 누르지 말라는 새끼들 땀										
11	요새 TV	여주일 프라	주일 프라이타임 예능 프로그램 고정이면 예초에 급 연예인들인데 호텔에서 삼시 세끼 먹어도 되는 양반들이										
12	남고 재직	장작을 주	장작을 주네 폐미교육 증거자료한국도 미국 일본처럼 학부모회 권한 확대할 필요성이 증 있다고 생각함 별 짓										
13	야근하다	소설 쓰네	소설 쓰네 이것만 하고 가면 안 되나요 지금 못하면 내일 쫓 되게 생겼습니다 유력유력일 단 임원인 니가 야근										
14	신남성연	정부에서	정부에서 삭제 요청 들어와서 채널 일주일 정치 수익 정치당했다고 함 나도 과거 보면 저 사람들 별로 긴 한데										
15	한강 사망	두 시부터	두 시부터 집회 시작 근데 주최자들 런 해서 누가 주최자인지도 모르는 상황자기네들 편이라 생각했던 그 알은										
16	이름 되면	파이함방	파이 함방 몇 시간 도 안 되서 파이 함방래 나락이었지만 이후리 함방 몇 시간 뒤 니가르 폭로전 발발 세										
17	99 걸러야	할 특목 특	소신 발언 하자면 저렇게 불러놓고 마음씨는 착 한 노인네들도 있음거스프링 내 팔 하나 찢리고 겨우 이 김점										
18	종이 만화	귀찮이 일	귀찮이 얼마나 흥행한 지 모르고 쓴 건가 웹툰이 잘나가긴 하지만 누가 보면 다 망한 줄 알겠음 심지어 예시에										
19	강철부대	논란은 있	논란은 있었지만 군 관련 전문가라 솔직히 군 관련해서는 이근 이지 뒹어놓고 싸도 중알 지 멋대로 뛰는 가스										

전체 형태소 개수:
18381
전체 명사 개수:
6440

Out [23] :

	단어	빈도수
0	사람	46
1	여자	41
2	아이유	38
3	남자	34
4	동생	24
5	이근	22
6	생각	22
7	누나	21
8	이랑	20
9	정도	20

- 본문 등 내용의 크롤링을 진행하다가, 차단되었다. 현재 다양한 우회 방법에 대해 자료조사를 하는 중이다.

커뮤니티 크롤링

02. 클리앙



```
result_path = './result.xlsx'
result_word_path = './result_word.xlsx'

if __name__ == "__main__":
    try:
        make_excel(result_path)
        print('엑셀 파일 생성.')
    except PermissionError:
        print('엑셀 파일 생성 불가. 열려 있는 엑셀 파일을 종료해주세요.')
        sys.exit(1)
    print('=====')

    start_idx = 1
    with tqdm(total=100, ncols=80, desc='파일 생성 중') as pbar:
        for item in get_data():
            start_idx = append_excel(result_path, item, start_idx)
            pbar.update(1)

    print('=====')
    print('작업 완료.')
    input()

    try:
        make_excel2(result_word_path)
        print('엑셀 파일 생성.')
```

33	쫄지,마,항상,응원,합니다,,	
34	응원,합니다	
35	항상,응원,합니다	
36	응원,합니다	
37	응원,합니다	
38	응원,합니다	
39	2 장인,께서,영,면,하셨습니다,...	7년,전,아버지,소천,하셨을,때,도,클리앙,에서,많은,위로,를,받았
40	삼가,고인,의,명복,을,빕니다	
41	삼가,고인,의,명복,을,빕니다,/,Vollago	
42	삼가,고인,의,명복,을,빕니다	
43	삼가,고인,의,명복,을,빕니다	
44	평안,에,머우시길,빕니다	
45	삼가,고인,의,명복,을,빕니다	
46	글,에서,장인어른,에,대한,존경,과,사랑,이,느껴지네요,고인,의,명복,을,빕니다	
47	삼가고인의,명복,을,빕니다	
48	삼가,고인,의,명복,을,빕니다	
49	고인,이,남긴,삶,의,향기,를,느낄,수,있는,좋은,부고,글,인,것,같습니,삼가,고인,의,명복,을,빕니다	
50	고인,의,명복,을,빕니다,좋은,곳,으로,가시길,...	
51	삼가,고인,의,명복,을,빕니다	
52	이제,평안,을,누리,시기를,기원,합니다	
53	삼가,고인,의,명복,을,빕니다	

분': 9, '재판': 3, '주위': 3, '생각': 3, '청년': 3, '지금': 3, '후': 2, '연락': 2,
2, '참고': 2, '감사': 2, '후원': 2, '저': 2, '스피커': 2, '양성': 2, '등': 2, '예
청소년': 2, '채널': 2, '준비': 2, '힘': 2, '어제': 1, '다스': 1, '출연': 1, '이야기
버': 1, '서도': 1, '신경': 1, '본의': 1, '분란': 1, '수': 1, '방송': 1, '응원': 1,
': 1, '비용': 1, '시민단체': 1, '사용': 1, '고양이': 1, '뉴스': 1, '개월': 1, '구
1, '무엇': 1, '문화': 1, '예술': 1, '정책': 1, '발굴': 1, '소통': 1, '가장': 1,

- 제목, 본문, 댓글을 크롤링하고 각각을 형태소 단위로 정리해보았다.

차주에는

1. 선정한 커뮤니티 3곳에 대한 크롤러 구현 진행
 2. 크롤러 구현 시 차단을 막기 위한 우회 방법 조사 및 추가
 3. 형태소 분석기 사전에 단어 추가하여 전처리 단계에서 제외되지 않도록 수정
 4. 감성 사전에 대한 학습 및 자료조사
- 을 진행하고자 합니다.



감사합니다 :)
빅데이터미네이터